Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 186 (2015) 431 – 439

Procedia
Social and Behavioral Sciences

5th World Conference on Learning, Teaching and Educational Leadership, WCLTA 2014

Entropy to Measure Intra-Subject Test-Retest Reliability

Ali Baykal[a]*

[a]*Bahcesehir University, Istanbul, 34353, Turkiye*

Abstract

Operational definitions of reliability contradict its conceptual definition. Correlational techniques and internal consistency measures emphasize inter-subject reliability but omit intra-subject reliability. The proposed approach asserts that randomness is the error rather than assuming that the error is random. Using the entropy concept borrowed from information theory an index has been defined to quantify intra-subject reliability. The proposed index of intra-subject test-retest reliability was computed for 823 subjects in a 5 point Likert scale self-report personality inventory. The analysis of findings demonstrates that the proposed quantifier is appropriate to describe all of the possible response configurations. Independent intra-subject reliability indices will replace raw scores in computing inter-subject reliability coefficients. In addition it will be possible to report the reliability of the measurements in single-subject research.

Peer-review under responsibility of Academic World Education and Research Center

Keywords: Intra-subject reliability; inter-subject reliability; single-subject research; evaluation research

1. Introduction

Reliability in a behavioral assessment can be defined at three different levels:
Conceptual Level: This is the most fundamental level to start with. At this level reliability is associated with its synonyms, substitutes or alternative expressions. These are stability, consistency, repeatability, reproducibility etc. Theoretical Level: Theoretically reliability refers the degree to which the observed scores are free from random error. Obviously, theoretical definition ignores the constant and the systematic errors at the very beginning. Since it

* Ali Baykal. Tel.: +090-212-3815169; fax: +090-212-381 0025.
E-mail address: ali.baykal@bahcesehir.edu.tr

Peer-review under responsibility of Academic World Education and Research Center
doi:10.1016/j.sbspro.2015.04.154

rests upon the response counts, it implicitly assumes at least interval level measurement that requires an arbitrary zero point and equidistant, (i.e. uniform, homogeneous) units. The content of a test describes and delineates its zero point.

The unit of all scores in the globe is called "point". But there are no two identical "points" in the entire world. The points assigned by the same scorer within the same test are all different in size. Uniformity of units in scoring is just an enchanted assumption. In fact, the theoretical definition of reliability ends up with an algebraically unsolvable equation with two unknowns: The variance of random error component in scores and the reliability itself.

Practical Level: Instead of inventing ways to estimate the amount of random error involved in a measurement, a myriad of operational definitions and formulas have been devised to report reliability in testing (Gulliksen, 1950; Guilford, 1965; Cronbach, 1975; Thompson, 2012). "Methods of reliability" may differ according to the "types of reliability" which are as follows:

1.1. Key Reliability

Key reliability refers to the consistency keyed responses prepared by the expert(s) at different independent trials. It is an essential pre-requisite for all the other types of reliability. In the assessment of divergent abilities, key reliability becomes an irrelevant or quite complex concept. It is mostly relevant to scoring the convergent dimensions in achievement and aptitude testing. In Likert scales, the consistency of the directions (plus or minus) of items set by different experts must be considered as key reliability. It can easily be assessed by the ratio between the total number of agreements on the response options for which credits will be given and the total number of items in the instrument. There is no room for randomness in setting the keys of a set of items in a test, therefore it has to be ensured by iterations that it is equal to unity.

1.2. Intra Scorer Reliability

When a test is given, there must be at least one source of authority to judge or rate the observed responses with respect to the expected ones. The degree to which the judgments given about the same set of responses at different trials are consistent is called intra subject reliability. Intra scorer reliability has nothing to do with honesty of the individual. Halo effect, fatigue, prejudice, personal bias and the like are the factors which may jeopardize the degree of reproducibility of scores assigned for the same set of responses by the same scorer. Measures to be taken against these dangers are not as practical as they are obvious. Intra subject reliability tends to be high in scoring choice type of items, and it certainly tends to decrease in free format of items. Virtually, the total scores comprise the benchmark for the comparison of different scores assigned by the same scorer. However, there is a possibility of getting the same composite score as a sum of the components scored inconsistently at different trials.

1.3. Inter Scorer Reliability

Reproducibility of scores is not always something desirable. Some persistent misconceptions can also be repeated by a single scorer. In some large scale open ended exams, there has to be a large number of scorers in order to grade the participants promptly. These are just a few reasons to employ more than one single scorer in an assessment program. The degree to which the scores of subjects can be obtained by different scorers independently is called inter-scorer reliability. The implicit assumption is that the average random error will approximate to zero when the same performance is scored infinitely many times by independent raters. Due to apparent physical, economic and social constraints in many cases, two or three independent scorers might be found to be sufficient in practice.

1.4. Intra Subject Reliability

As can be easily extrapolated from the previous definitions, intra subject reliability refers to the reproducibility of the identical responses (answers) to a variety of stimuli (items) by a single subject in two or more trials. Of course, this attribute is relevant to relatively permanent convergent skills. In the assessment of divergent skills, response flexibility overrides response stability which implies the number of repeated responses (answers) to a given number

of repeated (questions) in two or more trials. Total number of "right" answers (total score of an individual) is a biased demonstration of intra subject reliability. Same total score can be obtained from a variety of different response patterns. For instance, one can get 5 points out of 252 different ways in a test of 10 True/False items. Evidently persistence of misconceptions should also be considered within the scope of intra subject reliability. Systematic failure has to be discriminated from disorganized random response patterns. Gambling (multiple option random guessing experience) is the most relevant example for random response pattern. Statistically, the average chance success is equal to the number of options in each trial divided by the number of total trials. Getting 20 points out of 100 multiple choice items is the most probable (frequently observable) chance success. Getting zero point requires more substantial effort than getting 20 points in this event. To sum up, intra subject reliability calls for an effort to study response patterns rather than simple arithmetic operations on total scores.

1.5. Inter Subject Reliability

Unless otherwise specified inter subject reliability is meant in reliability analyses in behavioral testing. As had been told at the very beginning, measurement must be repeated infinitely many times to get rid of "random error". Reliability indicates the degree to which the test measures the same thing time after time and item after item. Consistency over time and items are basic to the concept of reliability (Tuckman, 1975). Instead of assessment of random error, five major operational approaches have been developed to assess inter subject reliability:

- Test-retest : Correlation between the scores obtained from two observations, measures stability
- Test-equivalent test : Correlation between the scores obtained from two observations, measures equivalence
- Split halves (parts) : Correlation between the scores obtained from two observation, measures consistency
- Internal consistency: Kuder-Richardson, Spearman-Brown, Gulliksen, Flanagan, Cronbach Alpha etc.

All of these operational definitions of reliability contradict its theoretical definition. In theory, reliability varies between nil (zero) and perfect (unity). Correlation coefficient, however, is invented to cover the range of relationships between minus one and plus one i.e., which are perfectly opposite in direction and perfectly aligned respectively. Negative reliability is indefinite within the theoretical context. Time interval between test and retest trials are subject to well-known threats such as maturation, subject loss, testing effects, learning and forgetting etc.

The Split-half method is essentially a correlational technique. Hence it is subject to all of the shortcomings of the correlational approach: First, split half reliability can take minus values which are theoretically undefined. Second, split-halves reliability will tend to be indefinite when variability among students ceases to exist on either or both halves of the test. Third, the magnitude of reliability is altered when there appear to be significantly different groups combined as subjects in the test. It is likely to observe non-reliability for the whole group although the test is significantly reliable for both of the groups separately. It is also possible to overestimate the reliability for the whole group although two combined groups of students perform inconsistently on two different halves of the test. Last, the presence of a few extreme scores might inflate the correlation coefficient overwhelmingly (McCall, 1975).

Internal consistency measures of reliability emphasize inter-subject reliability, but omit intra-subject reliability. A reliable subject may not be distinguished from non-reliable test-mates, even when the interaction among individuals has been put under restraint during the test. Standard error of measurement is constant for all subjects just because of the operational definition reliability. Even hypothetically, there is no practical explanation for having a constant standard error for a diverse range of observed scores. On the other hand; in case of having negative correlation in a test-retest method or between the split halves, standard error of measurement will be a complex (imaginary) number.

In order to obtain high reliability for the test, the optimum value for item difficulty is 0.5, and the optimum level of the average score is 50% of the total number of items in the test. These optimum values, which are suggested by Kuder-Richardson formulas, contradict the concept of construct validity. Instead of making what is being measured relevant to what is intended to be measured, the test maker is advised to control the difficulty of the test and that of the items in an artificial way.

All of the operational methods for reliability depend on individual differences. The reliability is undefined when inter-subject variability ceases to exist. Variance of scores will vanish in the absence of learning; when mastery learning is achieved; and when the achievement differences among students are leveled down at any point between these two extremes. In a perfectly efficient curriculum, however, the achievement of students must be zero in the

average at entry level, and must progress to full mastery when instruction terminates. If these desiderata occur, the reliabilities of pretest and of posttest will not be known.

1.6. Rasch Approach: Reliable Person, Reliable Item

Rasch asserted that the reaction of a subject to a problem is not deterministic but probabilistic. The behavior of a subject can be described in terms of the probability that the respondent accomplishes the given task. Also the probability of correct response depends on the respondent's ability (θ) and the difficulty of the question (b). His mathematical formulation as to the probability of getting an item right or wrong is a conjoint function of the ability of the subject and the difficulty of the item both of which are assumed to be independent of each other. Simply probability of solving a problem increases as the ability of the subject increases. On the other hand; probability of success decreases as the difficulty of item increases.

As to the reliability of both subjects and items, the percentage of reproducible observed responses matters. "Person reliability" is equivalent to the traditional "test" reliability (Gracia, 2005; Linacre, 2005). To increase subject reliability, extremely low and high level ability subjects must be included in the test with a large number of items. The item reliability in Rasch model depends on how different the items are in terms of their difficulty. Reliability for subjects and items ranges from 0 to 1. The closer the reliability is to 1, the less the variability of the measurement can be attributed to measurement error. The index of separation is another measure of the fit of the data to the Rasch model. Separation refers to the spread of person positions or item positions along the variable measured If item separation is equal or lower than unity, items do not construct a definite dimension. Similarly, subject separation indices smaller than unity imply that the scale cannot discriminate the respondents reliably. The lower limit for both item and subject separation index is 1.0. To be able to conclude that items have sufficient breadth and subjects are discriminated well enough, both of the separation indices must be higher than 1.0.

The first and the most frequent criticism is the assumption of unidimensionality which asserts that the items must relate only to one principal construct (Panayides&Robinson&Tymms, 2011). There are so many sub-dimensions in an achievement test just because of the content and taxonomical levels. Most of the attitudes are multi-trait characteristics. Even Hambleton (1993; 150), who is one of the leading proponents of Rasch, accepts that "the unidimensionality assumption cannot strictly come true because there are always other cognitive, personality and test-taking factors that affect test performance, "at least to some extent". In fact, construct validity analysis of items reveal the fact that items are multidimensional to a great extent. Verbal abilities, for instance, are integral parts of almost every kind of testing. Chance, speed in perception and response, intrinsic and/or extrinsic motivation, attention, previous knowledge in topic and interest in test content etc. are factors which are apparently present in the measurement of other constructs.

It is quite ironic that, in a probabilistic model, ability of subject is constant for all items, and the difficulty of an item is constant for all subjects. It is quite difficult to take this assumption for granted when there are so many real life examples on the contrary. Difficulty order of items is different for all subjects. It is very unlikely that response patterns of subjects are similar even among the ones at the same ability level. Invariance in Rasch model is a mystical expectation corresponding to "true score" in classical test theory.

Rasch measurement requires equally discriminating items to satisfy the uniformity assumption. Hence unequal item discriminations are regarded as item malfunctioning or distortion of measurement. The aim of measurement is not to fit the data to the model but to enable a model to describe the reality. If subjects are the same along the construct being measured there needs to be no discrimination at all. If subjects are different along the dimension then items must enable the observer to discriminate.

Conclusively total test response or particular item responses can only be expressed in probabilistic terms. Any description in terms of probabilities is inevitably imprecise. Therefore, even the Rasch model cannot portray any construct free from error.

2. Proposal: Reliability is the extent to which response departs away from randomness

All of the above as well as other shortcomings mentioned above stems from the assumption that the error varies randomly within the observed achievement scores. Aside from the fact that the assumption does not hold in many

situations, its implications are not convenient for practical purposes. Error may or not be random but randomness is definitely an error.

2.1. Assumption

The approach proposed in this paper is based upon the view that the randomness is the error, instead of assuming that the error is random. Hence, a measure of randomness is needed. In physics the entropy, in information theory the average uncertainty are measures of the randomness, chance, and aimlessness. The tendency in a system to proceed towards a state of greater disorder is expressed by the concept of entropy. When the system becomes more and more disorganized, one is less informed than before. In testing situations there are at least two courses of action, one of which is keyed to be the right choice. In other words, the test maker creates some uncertainty situations in order to see if the respondent is sure about what to do. Students are required to break the hidden codes in such a way as to remove the uncertainty. Therefore, learning can be measured in terms of the departure from complete uncertainty which is synonymous with ignorance (McGill, 1954; Garner&McGill, 1956; Attneave, 1959; Omurtak, 1972).

2.2. Frequency distribution of responses of a single-subject in a test-retest experiment

The truth and the error can be defined with respect to the purpose of the experiment (Turgut, 1975;18-19). In this study, the error is not a random component within the observed score. It is defined as uncertainty in the distribution of retest responses which cannot be explained by the choices in the pretest. Intra-subject reliability can be measured starting from the point where the retest responses are completely independent of the test responses i.e. in terms of the decreasing uncertainty departing from the maximum possible depending upon the item format. In a multiple choice test or in a Likert scale, there are "a+1" alternative responses. "a" is the number of alternatives which can be chosen plus 1 refers to all the other possibilities put together e.g. omissions, double choices etc. Test-retest responses of every single individual can be plotted on a frequency matrix as shown in Table. 1 below.

Table 1. Frequency matrix for test-retest responses

	A	B	C	D	E	F	Total $f(x)$
A	$f(x_1,y_1)$						
B							
C			$f(x_3,y_3)$				
D							
E							$f(y_5)$
F						$f(x_6,y_6)$	
Total f(y)				$f(x_4)$			Grand Total

In this table, $f(x_i,y_j)$ stands for the frequency of responses observed for the i^{th} option on the posttest corresponding to the j^{th} option on the pretest. The marginal total $f(x)$ represents the frequency of optional choices made by the subject on the pretest. Similarly, $f(y)$ is the frequency of choices for each option on the posttest. These are defined by the formulas (1) and (2) respectively.

$$\text{Total frequency of responses for the option "i" in pretest}: f(x_i)=\sum_{j=1}^{a+1} f(x_i,y_j) \quad (i=1,2,\ldots,a+1) \tag{1}$$

$$\text{Total frequency of responses for the option "j" in retest}: f(y_j)=\sum_{i=1}^{a+1} f(x_i,y_j) \quad (j=1,2,\ldots,a+1) \tag{2}$$

2.3. Shannon's entropy formulas to measure uncertainty in information exchange

Shannon (1949) defined entropy as a quantitative measure of "noise" in a two way communication experiment. Here, pretest responses observed for A, B, …, F choices correspond to signals sent. Respective choices in retest correspond to the signals perceived. The joint entropy measures how much uncertainty is enclosed within the cross-tabulated pretest-posttest responses to K number of items along a+1 response options. The agreement between test-retest responses corresponds to mutual information which quantifies the reproducibility of test-retest responses, but not the total scores obtained. If the test and retest responses are completely independent the uncertainty will be maximum which denotes absolute absence of reliability. When there is a perfect match between the test and retest responses, the mutual information will be maximum that implies perfect reliability.

$$\text{The uncertainty observed in pretest responses}: \quad H(X) = -\sum_{i=1}^{a+1} p(x_i) \ln p(x_i) \tag{3}$$

$$\text{The uncertainty observed in retest responses}: \quad H(Y) = -\sum_{j=1}^{a+1} p(y_j) \ln p(y_j) \tag{4}$$

$$\text{The joint uncertainty between test} - \text{retest responses}: \quad H(X,Y) = -\sum_{i=1}^{a+1}\sum_{j=1}^{a+1} p(x_i, y_j) \ln p(x_i, y_j) \tag{5}$$

$$\text{Probability equivalent of any}(z) \text{ observed frequency in test and \ or in retest}: \quad p(z) = \frac{f(z)}{K} \tag{6}$$

Where K is the total number of items compared between test and retest.

$$\text{Measure of consistency between test and retest responses}: \quad I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{7}$$

These formulas can be used to describe intra-subject reliability of a single subject (Baykal, 1980).

$$\text{Index of intra}(\text{within}) \text{ subject reliability in a test / retest experiment}: \quad \rho_W = \frac{I(X,Y)}{H(X,Y)} \qquad 0 < \rho_W < 1 \tag{8}$$

3. Practice: Inter subject reliabilities of 823 subjects in TAT

3.1. Instrument and its validity

TAT (TANINMA ALGISI TUTANAGI) ~ Inventory of Perceptions of Others as Perceived by Self) is a five-point Likert scale self-reported personality inventory of 120 items in Turkish. TAT consists of six sub-constructs:

- Dynamism: Active, energetic, enthusiastic, self-initiative, rootless, nomadic;
- Achievement motivation: Goal oriented, drive to succeed, stimulated, provocative;
- Sociability: Sociable, representable, friendly, warm, prone, extrovert;
- Flexibility: Adaptable, compliant, malleable, accommodating;
- Assertiveness: Confident, insistent, persistent, persuasive;
- Risk taking: Risk taker, investor, entrepreneur,

These sub-constructs were demanded by the top administrators of a newly established bank in 1992. Verdict statements, which are likely to reveal personality traits were written by the author who was inspired of Eysenck (Eysenck&Wilson, 1975) CPI (California Personality Inventory) and to some extent MMPI (Minnesota Multitrait Personality Inventory). The relevancy of these statements to the sub-constructs concerned were discussed and analyzed in "hermeneutic" sessions held together with the author and the authorities in Human Resources Department. The content validity is limited to 6 sub-constructs although there are 20 in CPI and more in MMPI. Then, the items were given to some experts in education, business administration and psychology. As to the distribution of items to sub-constructs strong agreement have been ensured among the experts through iterative

corrections. A series of factor analyses, which have been done after large group participations, yielded reinforcing evidence for the appropriateness of TAT.

TAT and its shortened versions were administered to more than 20000 job applicants for the employment in well-known companies in Turkey within the period of 1992-2002. It was one of the components of a battery which was used as a screening device. First one was a general aptitude test of 80 multiple choice items. The second one was a test of factual information about plastic arts, music, politics, sciences, literature etc. First stage of assessment can be considered as a high-stake assessment because applicants were ranked according to their composite scores obtained from these tests. Although the weight of TAT was almost nil the applicants had always been told that they might have been inquired about their responses in TAT during the interview.

35 different firms or companies applied TAT twice or more. There is no statistical evidence for its predictive validity. However there are some cues and clues that the inclusion error is not so destructive, otherwise it wouldn't have been demanded by so many companies repeatedly over the years. In fact, the exclusion error of selection tests cannot ever be known due to ethical and practical reasons.

Like any other test of "personality traits", the face validity of TAT is and will always be questionable.

3.2. Reliability of TAT

No matter to what extent a test measures what it purports to measure it must measure what it is supposed to measure consistently. Almost all kinds of methods yielded higher reliabilities than satisfactory for the sub-constructs and the whole. It is neither possible nor desirable to display their data. Table 2 summarizes the test and item statistics for 18650 participants in general and 823 subjects who participated twice.

Table 2. Some selected test and item statistics and reliability indicators of TAT.

Scale:	A-Odd	A-Even	A-Tot	Pre-Odd	Pre-Even	Pre-Tot	Re-Odd	Re-Even	Re-Tot
N of Items	60	60	120	60	60	120	60	60	120
Number of Options	5	5	5	5	5	5	5	5	5
N of Examinees	18650	18650	18650	823	823	823	823	823	823
Mean	2.505	2.604	2.555	2.483	2.589	2.513	2.394	2.498	2.421
Variance	0.156	0.106	0.112	0.192	0.171	0.177	0.218	0.179	0.193
Std. Dev.	0.395	0.325	0.335	0.439	0.413	0.421	0.467	0.423	0.440
Skew	0.176	0.034	0.046	0.032	0.141	0.035	0.196	0.303	0.234
Minimum	1.000	1.000	1.000	1.483	1.717	1.608	1.400	1.000	1.500
Maximum	5.000	5.000	5.000	3.643	3.765	3.427	3.520	3.765	3.425
Median	2.450	2.600	2.525	2.450	2.533	2.433	2.300	2.417	2.308
SEM	0.154	0.159	0.111	0.152	0.157	0.109	0.148	0.154	0.106
Mean Item-Tot.	0.320	0.261	0.273	0.352	0.324	0.337	0.379	0.337	0.356
Alpha	0.848	0.761	0.890	0.880	0.855	0.933	0.900	0.868	0.941
K-R 21	0.819	0.882	0.867	0.773	0.800	0.785	0.740	0.789	0.765
Spearman-Brown	-	-	0.839	-	-	0.904	-	-	0.910

The abbreviations in Table 2 represent the followings:

Table3. The abbreviations in Table 2

Acronym		Samples of participants and
A-Odd	:	Split half of TAT covering odd numbered items with 18650 participants
A-Even	:	Split half of even numbered items with 18650 participants
A-Tot	:	The whole TAT with 18650 participants
Pre-Odd	:	Split half of TAT covering odd numbered items in pretest with 823 participants
Pre-Even	:	Split half of TAT covering even numbered items in pretest with 823 participants
Pre-Tot	:	The whole TAT in pretest with 823 participants
Po-Odd	:	Split half of TAT covering odd numbered items in retest with 823 participants
Po-Even	:	Split half of TAT covering even numbered items in retest with 823 participants
Po-Tot	:	The whole TAT in retest with 823 participants

Test and item statistics in Table 2 were accepted as being optimum to proceed to use TAT in this study.

3.3. Procedure for the evaluation

Intra-subject test-retest reliabilities of these participants were computed in four different ways:

i. Total number of consistent responses between test and retest data were counted for all subjects.

ii. Intra subject reliability coefficients ($\rho_{w)}$ of all participants were computed as proposed in 2.3.

Random test and retest responses of 823 subjects were obtained for 120 items with 5 options. That was done by using the RANDBETWEEN(1;5) function in MSO EXCEL software.

iii. Total number of consistent responses between simulated test and retest data were counted for all subjects.

iv. Intra subject reliability coefficients ($\rho_{w)}$ of all participants were computed for randomized responses.

Table 4. is a summary of the correlations between the pairs of intra-subject measures:

Table 4. Correlations between intra-subject consistency measures for the test-retest responses of 823 subjects

Measure of response consistency *	CODE	ENT	AGR	RENT
Entropy based intra-subject reliability in test-retest responses (TAT)	ENT	1		
Percentage of agreement between test-retest responses (TAT)	AGR	.905**	1	
Entropy based intra-subject reliability for randomized test-retest responses	RENT	-0.001	-0.017	1
Percentage of agreement between randomized test-retest responses	RAGR	0.003	-0.003	-0.030

* N=823 for all **. Correlation is significant at the 0.01 level (2-tailed).

The findings in Table 4 are compatible with the expectations. The basic conjecture of correlation is apparent that random variables are very unlikely to yield significant correlations with others. High positive correlation between REBR and RPAG may seem to be autocorrelation, but it is not. There is a possibility for REBR to be equal to unity while there is no one-to-one equivalence but one to one correspondence. One of the subjects, for instance, might have chosen Bs in retest corresponding to all As the in pre-test, Cs in pre-test might have been replaced with Bs; so and so forth. On the other hand, there is a probability that all or some of the agreements might have been due to chance which is a perfectly real example for random events.

3.4. Progress

In spite of all these possibilities computing entropy based intra-subject reliability indices may not seem to be practical in comparison with simple counts of agreements. What remains is the intellectual value of theoretical coherence. There are some other entropy based indicators which can be used to quantify item and subject characteristics (Maccia, 1963; Hintikka&Suppes, 1970; Guiaşu, 1977).

Almost all measures of traditional assessment depend on variance of scores. Therefore, they cannot be used in individual testing programs. Lack of variation among students may occur at any point between an absence of learning and perfect mastery. In a condition that two groups of students are given a pre-test measuring the objectives of a successful curriculum the pre-test results must display a complete incompetency for all students if the instruction is perfectly needed. Suppose a new strategy for mastery is studied in one of the groups while the other is treated conventionally as a control group. Assume the proposed strategy lead all subjects to perfect mastery, and the conventional strategy produce a normal distribution along achievement as would be predicted. In such an observation, only the reliability of the post-test given to the control group can be reported. The reliability of the post-test given to the experimental group as well as of the pre-tests given to both groups will be unknown. In fact, no parametric statistical inference can be made in such an experiment.

Since the information conveyed through wrong responses is also being used, the predictive power of items can be determined more accurately. This has applications in testing for selection. Equivalency of items can be measured more precisely. Thus, parallel forms of tests can easily be prepared when needed.

To sum up, the proposed approach has been found to be convenient for both criterion-referenced and norm-

referenced evaluation practices (Popham, 1971; Meskauskas, 1976). Further empirical studies will be carried out to display the uses of the approach.

Acknowledgements

This paper is a by-product of the research project titled *Using Entropy Indices in the Analysis of Test Data* encoded as 09D6038. It has been supported by Bogazici University Scientific Research Projects Fund. The author wishes to express his gratitude to Administrative Coordination Office for Research Projects for their helping hand.

References

Attneave, F. (1959). *Applications of information theory to psychology*. New York: Henry Holt and Company.

Baykal, A. (1981). Entropy as a measure of error in achievement testing. *Bogazici Journal of Education. Vol. VIII*-IX; 1980-81.

Cronbach, L J. (1970). *Essentials of psychological testing.* Third Edition. New York: Harper and Row.

Eysenck, H.J., Wilson, G. (1975). Know your own personality..Penguin Books, Middlesex, England.

Garner, W. R., WJ. McGill. (September 1956). The relation between information and variance analyses. *Psychometrika*. 21-3, 219-228.

Gracia, S. (2005). *Analyzing CSR implementation with the Rasch Model*. Rhode Island College. (http://digitalcommons.ric.edu/cgi/viewcontent.cgi?article=1269&context=facultypublications)

Guiaşu, S. (1977). *Information theory with applications*. New York: McGraw-Hill.

Guilford J.P. (1965). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley and Sons Inc.

Hambleton, R. K. (1993). Principles and selected applications of Item Response Theory, in: R.L Linn, (Ed), *Educational Measurement*. (3rd ed.) 13-104. Phoenix: Oryx Press.

Hintikka, J., P. Suppes, (Eds.). *Information and inference*. Dordrecht, Holland: D. Reidel Pub. Co., 1970.

Linacre, J.M. !1997). KR-20 / Cronbach Alpha or Rasch Person Reliability: Which tells the "Truth"? *Rasch Measurement Transactions*. (http://www.rasch.org/rmt/rmt113l.htm)

Maccia, E.S. (1963). An Educational Theory Model: Information Theory. *Occasional Paper 63-141.* Ohio: Bureau of Educational Research and Service,

McCall, R.B. (1975). *Fundamental statistics for psychology. (Second Edition)*. New York: Harcourt Brace Jovanovich, Inc.

McGill, W.J. (June 1954). Multivariate information transmission. *Psychometrika.* 19:2, 97-116.

Meskauskas, J.A. (Winter 1976). Evaluation models for criterion-referenced testing, *Review of Educational Research.* 46:1, 133-158.

Omurtak, Y. (1972). Toward an operational definition of education and its measurement in non-monetary terms: education and entropy, Seminar Notes, Ankara: 1972.

Panayides, P., Robinson, C., Tymms, P. (2011). The Assessment revolution that has passed England by: Rasch Measurement. *Durham Research Online*; Durham University. (http://dro.dur.ac.uk/6405/1/6405.pdf?DDD29+ded4ss+d67a9y)

Popham,W.J. (Ed.). (1971). *Criterion-referenced measurement*. New Jersey: Educational Technology Publications.

Shannon, C.E., Weaver, W. (1949). *The mathematical theory of communication*. Urbana: The University of Illinois Press.

Thompson, B. (2012). Score reliability: Contemporary thinking on reliability issues. Sage publications.

Tuckman, B.W. (1975). *Measuring educational outcomes*. New York: Harcourt Brace Jovanovich.

Turgut, M. F. (June-December 1975). Theories of error and estimating the errors of measurement. *Hacettepe Bulletin of Social Sciences and Humanities*. 7:1-2, 1-20.